A Mathematical Theory of Communication

By C. E. SHANNON

(*Concluded from July 1948 issue*)

PART III: MATHEMATICAL PRELIMINARIES

In this final installment of the paper we consider the case where the signals or the messages or both are continuously variable, in contrast with the discrete nature assumed until now. To a considerable extent the continuous case can be obtained through a limiting process from the discrete case by dividing the continuum of messages and signals into a large but finite number of small regions and calculating the various parameters involved on a discrete basis. As the size of the regions is decreased these parameters in general approach as limits the proper values for the continuous case. There are, however, a few new effects that appear and also a general change of emphasis in the direction of specialization of the general results to particular cases.

We will not attempt, in the continuous case, to obtain our results with the greatest generality, or with the extreme rigor of pure mathematics, since this would involve a great deal of abstract measure theory and would obscure the main thread of the analysis. A preliminary study, however, indicates that the theory can be formulated in a completely axiomatic and rigorous manner which includes both the continuous and discrete cases and many others. The occasional liberties taken with limiting processes in the present analysis can be justified in all cases of practical interest.

18. Sets and Ensembles of Functions

We shall have to deal in the continuous case with sets of functions and ensembles of functions. A set of functions, as the name implies, is merely a class or collection of functions, generally of one variable, time. It can be specified by giving an explicit representation of the various functions in the set, or implicitly by giving a property which functions in the set possess and others do not. Some examples are:

1. The set of functions:

$$f_\theta(t) = \sin(t + \theta).$$

Each particular value of θ determines a particular function in the set.

2. The set of all functions of time containing no frequencies over W cycles per second.
3. The set of all functions limited in band to W and in amplitude to A.
4. The set of all English speech signals as functions of time.

An *ensemble* of functions is a set of functions together with a probability measure whereby we may determine the probability of a function in the set having certain properties.[1] For example with the set,

$$f_\theta(t) = \sin (t + \theta),$$

we may give a probability distribution for θ, $P(\theta)$. The set then becomes an ensemble.

Some further examples of ensembles of functions are:

1. A finite set of functions $f_k(t)$ $(k = 1, 2, \cdots, n)$ with the probability of f_k being p_k.
2. A finite dimensional family of functions

$$f(\alpha_1, \alpha_2, \cdots, \alpha_n ; t)$$

with a probability distribution for the parameters α_i:

$$p(\alpha_1, \cdots, \alpha_n)$$

For example we could consider the ensemble defined by

$$f(a_1, \cdots, a_n, \theta_1, \cdots, \theta_n ; t) = \sum_{n=1}^{n} a_n \sin n(\omega t + \theta_n)$$

with the amplitudes a_i distributed normally and independently, and the phrases θ_i distributed uniformly (from 0 to 2π) and independently.

3. The ensemble

$$f(a_i, t) = \sum_{n=-\infty}^{+\infty} a_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

with the a_i normal and independent all with the same standard deviation $\sqrt{N}$. This is a representation of "white" noise, band-limited to the band from 0 to W cycles per second and with average power N.[2]

[1] In mathematical terminology the functions belong to a measure space whose total measure is unity.

[2] This representation can be used as a definition of band limited white noise. It has certain advantages in that it involves fewer limiting operations than do definitions that have been used in the past. The name "white noise," already firmly intrenched in the literature, is perhaps somewhat unfortunate. In optics white light means either any continuous spectrum as contrasted with a point spectrum, or a spectrum which is flat with *wavelength* (which is not the same as a spectrum flat with frequency).

4. Let points be distributed on the t axis according to a Poisson distribution. At each selected point the function $f(t)$ is placed and the different functions added, giving the ensemble

$$\sum_{k=-\infty}^{\infty} f(t + t_k)$$

where the t_k are the points of the Poisson distribution. This ensemble can be considered as a type of impulse or shot noise where all the impulses are identical.

5. The set of English speech functions with the probability measure given by the frequency of occurrence in ordinary use.

An ensemble of functions $f_\alpha(t)$ is *stationary* if the same ensemble results when all functions are shifted any fixed amount in time. The ensemble

$$f_\theta(t) = \sin(t + \theta)$$

is stationary if θ distributed uniformly from 0 to 2π. If we shift each function by t_1 we obtain

$$\begin{aligned} f_\theta(t + t_1) &= \sin(t + t_1 + \theta) \\ &= \sin(t + \varphi) \end{aligned}$$

with φ distributed uniformly from 0 to 2π. Each function has changed but the ensemble as a whole is invariant under the translation. The other examples given above are also stationary.

An ensemble is *ergodic* if it is stationary, and there is no subset of the functions in the set with a probability different from 0 and 1 which is stationary. The ensemble

$$\sin(t + \theta)$$

is ergodic. No subset of these functions of probability $\neq 0, 1$ is transformed into itself under all time translations. On the other hand the ensemble

$$a \sin(t + \theta)$$

with a distributed normally and θ uniform is stationary but not ergodic. The subset of these functions with a between 0 and 1 for example is stationary.

Of the examples given, 3 and 4 are ergodic, and 5 may perhaps be considered so. If an ensemble is ergodic we may say roughly that each function in the set is typical of the ensemble. More precisely it is known that with an ergodic ensemble an average of any statistic over the ensemble is equal (with probability 1) to an average over all the time translations of a

particular function in the set.[3] Roughly speaking, each function can be expected, as time progresses, to go through, with the proper frequency, all the convolutions of any of the functions in the set.

Just as we may perform various operations on numbers or functions to obtain new numbers or functions, we can perform operations on ensembles to obtain new ensembles. Suppose, for example, we have an ensemble of functions $f_\alpha(t)$ and an operator T which gives for each function $f_\alpha(t)$ a result $g_\alpha(t)$:

$$g_\alpha(t) = Tf_\alpha(t)$$

Probability measure is defined for the set $g_\alpha(t)$ by means of that for the set $f_\alpha(t)$. The probability of a certain subset of the $g_\alpha(t)$ functions is equal to that of the subset of the $f_\alpha(t)$ functions which produce members of the given subset of g functions under the operation T. Physically this corresponds to passing the ensemble through some device, for example, a filter, a rectifier or a modulator. The output functions of the device form the ensemble $g_\alpha(t)$.

A device or operator T will be called invariant if shifting the input merely shifts the output, i.e., if

$$g_\alpha(t) = Tf_\alpha(t)$$

implies

$$g_\alpha(t + t_1) = Tf_\alpha(t + t_1)$$

for all $f_\alpha(t)$ and all t_1. It is easily shown (see appendix 1) that if T is invariant and the input ensemble is stationary then the output ensemble is stationary. Likewise if the input is ergodic the output will also be ergodic.

A filter or a rectifier is invariant under all time translations. The operation of modulation is not since the carrier phase gives a certain time structure. However, modulation is invariant under all translations which are multiples of the period of the carrier.

Wiener has pointed out the intimate relation between the invariance of physical devices under time translations and Fourier theory.[4] He has

[3] This is the famous ergodic theorem or rather one aspect of this theorem which was proved is somewhat different formulations by Birkhoff, von Neumann, and Koopman, and subsequently generalized by Wiener, Hopf, Hurewicz and others. The literature on ergodic theory is quite extensive and the reader is referred to the papers of these writers for precise and general formulations; e.g., E. Hopf "Ergodentheorie" Ergebnisse der Mathematic und ihrer Grenzgebiete, Vol. 5, "On Causality Statistics and Probability" Journal of Mathematics and Physics, Vol. XIII, No. 1, 1934; N. Weiner "The Ergodic Theorem" Duke Mathematical Journal, Vol. 5, 1939.

[4] Communication theory is heavily indebted to Wiener for much of its basic philosophy and theory. His classic NDRC report "The Interpolation, Extrapolation, and Smoothing of Stationary Time Series," to appear soon in book form, contains the first clear-cut formulation of communication theory as a statistical problem, the study of operations

shown, in fact, that if a device is linear as well as invariant Fourier analysis is then the appropriate mathematical tool for dealing with the problem.

An ensemble of functions is the appropriate mathematical representation of the messages produced by a continuous source (for example speech), of the signals produced by a transmitter, and of the perturbing noise. Communication theory is properly concerned, as has been emphasized by Wiener, not with operations on particular functions, but with operations on ensembles of functions. A communication system is designed not for a particular speech function and still less for a sine wave, but for the ensemble of speech functions.

19. Band Limited Ensembles of Functions

If a function of time $f(t)$ is limited to the band from 0 to W cycles per second it is completely determined by giving its ordinates at a series of discrete points spaced $\frac{1}{2W}$ seconds apart in the manner indicated by the following result.[5]

Theorem 13: Let $f(t)$ contain no frequencies over W.
Then

$$f(t) = \sum_{-\infty}^{\infty} X_n \frac{\sin \pi(2Wt - n)}{\pi(2Wt - n)}$$

where

$$X_n = f\left(\frac{n}{2W}\right).$$

In this expansion $f(t)$ is represented as a sum of orthogonal functions. The coefficients X_n of the various terms can be considered as coordinates in an infinite dimensional "function space." In this space each function corresponds to precisely one point and each point to one function.

A function can be considered to be substantially limited to a time T if all the ordinates X_n outside this interval of time are zero. In this case all but $2TW$ of the coordinates will be zero. Thus functions limited to a band W and duration T correspond to points in a space of $2TW$ dimensions.

A subset of the functions of band W and duration T corresponds to a region in this space. For example, the functions whose total energy is less

---

on time series. This work, although chiefly concerned with the linear prediction and filtering problem, is an important collateral reference in connection with the present paper. We may also refer here to Wiener's forthcoming book "Cybernetics" dealing with the general problems of communication and control.

[5] For a proof of this theorem and further discussion see the author's paper "Communication in the Presence of Noise" to be published in the *Proceedings of the Institute of Radio Engineers*.

than or equal to E correspond to points in a $2TW$ dimensional sphere with radius $r = \sqrt{2WE}$.

An *ensemble* of functions of limited duration and hand will be represented hy a prohahility distribution $p(x_1 \cdots x_n)$ in the corresponding n dimensional space. If the ensemhle is not limited in time we can consider the $2TW$ coordinates in a given interval T to represent suhstantially the part of the function in the interval T and the prohability distribution $p(x_1, \cdots, x_n)$ to give the statistical structure of the ensemhle for intervals of that duration.

20. Entropy of a Continuous Distribution

The entropy of a discrete set of prohahilities $p_1, \cdots p_n$ has been defined as:

$$H = -\sum p_i \log p_i .$$

In an analogous manner we define the entropy of a continuous distrihution with the density distrihution function $p(x)$ hy:

$$H = -\int_{-\infty}^{\infty} p(x) \log p(x)\, dx$$

With an n dimensional distrihution $p(x_1, \cdots, x_n)$ we have

$$H = -\int \cdots \int p(x_1 \cdots x_n) \log p(x_1, \cdots, x_n)\, dx_1 \cdots dx_n .$$

If we have two arguments x and y (which may themselves he multi-dimensional) the joint and conditional entropies of $p(x, y)$ are given hy

$$H(x, y) = -\int\int p(x, y) \log p(x, y)\, dx\, dy$$

and

$$H_x(y) = -\int\int p(x, y) \log \frac{p(x, y)}{p(x)}\, dx\, dy$$

$$H_y(x) = -\int\int p(x, y) \log \frac{p(x, y)}{p(y)}\, dx\, dy$$

where

$$p(x) = \int p(x, y)\, dy$$

$$p(y) = \int p(x, y)\, dx.$$

The entropy of continuous distrihutions have most (hut not all) of the properties of the discrete case. In particular we have the following:

1. If x is limited to a certain volume v in its space, then $H(x)$ is a maximum and equal to $\log v$ when $p(x)$ is constant $\left(\frac{1}{v}\right)$ in the volume.
2. With any two variables x, y we have

$$H(x, y) \leq H(x) + H(y)$$

with equality if (and only if) x and y are independent, i.e., $p(x, y) = p(x)\,p(y)$ (apart possibly from a set of points of probability zero).
3. Consider a generalized averaging operation of the following type:

$$p'(y) = \int a(x, y)p(x)\,dx$$

with

$$\int a(x, y)\,dx = \int a(x, y)\,dy = 1, \qquad a(x, y) \geq 0.$$

Then the entropy of the averaged distribution $p'(y)$ is equal to or greater than that of the original distribution $p(x)$.
4. We have

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x)$$

and

$$H_x(y) \leq H(y).$$

5. Let $p(x)$ be a one-dimensional distribution. The form of $p(x)$ giving a maximum entropy subject to the condition that the standard deviation of x be fixed at σ is gaussian. To show this we must maximize

$$H(x) = -\int p(x) \log p(x)\,dx$$

with

$$\sigma^2 = \int p(x)x^2\,dx \quad \text{and} \quad 1 = \int p(x)\,dx$$

as constraints. This requires, by the calculus of variations, maximizing

$$\int [-p(x) \log p(x) + \lambda p(x)x^2 + \mu p(x)]\,dx.$$

The condition for this is

$$-1 - \log p(x) + \lambda x^2 + \mu = 0$$

and consequently (adjusting the constants to satisfy the constraints)

$$p(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x^2/2\sigma^2)}.$$

Similarly in n dimensions, suppose the second order moments of $p(x_1, \cdots, x_n)$ are fixed at A_{ij}:

$$A_{ij} = \int \cdots \int x_i x_j p(x_1, \cdots, x_n)\, dx_1 \cdots dx_n .$$

Then the maximum entropy occurs (by a similar calculation) when $p(x_1, \cdots, x_n)$ is the n dimensional gaussian distribution with the second order moments A_{ij}.

6. The entropy of a one-dimensional gaussian distribution whose standard deviation is σ is given by

$$H(x) = \log \sqrt{2\pi e}\,\sigma .$$

This is calculated as follows:

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x^2/2\sigma^2)} \\ -\log p(x) &= \log \sqrt{2\pi}\,\sigma + \frac{x^2}{2\sigma^2} \\ H(x) &= -\int p(x) \log p(x)\, dx \\ &= \int p(x) \log \sqrt{2\pi}\,\sigma\, dx + \int p(x) \frac{x^2}{2\sigma^2} dx \\ &= \log \sqrt{2\pi}\,\sigma + \frac{\sigma^2}{2\sigma^2} \\ &= \log \sqrt{2\pi}\,\sigma + \log \sqrt{e} \\ &= \log \sqrt{2\pi e}\,\sigma . \end{aligned}$$

Similarly the n dimensional gaussian distribution with associated quadratic form a_{ij} is given by

$$p(x_1, \cdots, x_n) = \frac{|a_{ij}|^{\frac{1}{2}}}{(2\pi)^{n/2}} \exp\left(-\tfrac{1}{2}\Sigma a_{ij} X_i X_j\right)$$

and the entropy can be calculated as

$$H = \log (2\pi e)^{n/2} |a_{ij}|^{\frac{1}{2}}$$

where $|a_{ij}|$ is the determinant whose elements are a_{ij}.

7. If x is limited to a half line ($p(x) = 0$ for $x \leq 0$) and the first moment of x is fixed at a:

$$a = \int_0^\infty p(x) x\, dx,$$

then the maximum entropy occurs when

$$p(x) = \frac{1}{a} e^{-(x/a)}$$

and is equal to log ea.

8. There is one important difference between the continuous and discrete entropies. In the discrete case the entropy measures in an *absolute* way the randomness of the chance variable. In the continuous case the measurement is *relative to the coordinate system.* If we change coordinates the entropy will in general change. In fact if we change to coordinates $y_1 \cdots y_n$ the new entropy is given by

$$H(y) = \int \cdots \int p(x_1 \cdots x_n) J\left(\frac{x}{y}\right) \log p(x_1 \cdots x_n) J\left(\frac{x}{y}\right) dy_1 \cdots dy_n$$

where $J\left(\frac{x}{y}\right)$ is the Jacobian of the coordinate transformation. On expanding the logarithm and changing variables to $x_1 \cdots x_n$, we obtain:

$$H(y) = H(x) - \int \cdots \int p(x_1, \cdots, x_n) \log J\left(\frac{x}{y}\right) dx_1 \cdots dx_n .$$

Thus the new entropy is the old entropy less the expected logarithm of the Jacobian. In the continuous case the entropy can be considered a measure of randomness *relative to an assumed standard*, namely the coordinate system chosen with each small volume element $dx_1 \cdots dx_n$ given equal weight. When we change the coordinate system the entropy in the new system measures the randomness when equal volume elements $dy_1 \cdots dy_n$ in the new system are given equal weight.

In spite of this dependence on the coordinate system the entropy concept is as important in the continuous case as the discrete case. This is due to the fact that the derived concepts of information rate and channel capacity depend on the *difference* of two entropies and this difference *does not* depend on the coordinate frame, each of the two terms being changed by the same amount.

The entropy of a continuous distribution can be negative. The scale of measurements sets an arbitrary zero corresponding to a uniform distribution over a unit volume. A distribution which is more confined than this has less entropy and will be negative. The rates and capacities will, however, always be non-negative.

9. A particular case of changing coordinates is the linear transformation

$$y_j = \sum_i a_{ij} x_i .$$

In this case the Jacobian is simply the determinant $|a_{ij}|^{-1}$ and

$$H(y) = H(x) + \log |a_{ij}|.$$

In the case of a rotation of coordinates (or any measure preserving transformation) $J = 1$ and $H(y) = H(x)$.

21. Entropy of an Ensemble of Functions

Consider an ergodic ensemble of functions limited to a certain band of width W cycles per second. Let

$$p(x_1 \cdots x_n)$$

be the density distribution function for amplitudes $x_1 \cdots x_n$ at n successive sample points. We define the entropy of the ensemble per degree of freedom by

$$H' = -\operatorname*{Lim}_{n\to\infty} \frac{1}{n} \int \cdots \int p(x_1 \cdots x_n) \log p(x_1, \cdots, x_n)\, dx_1 \cdots dx_n .$$

We may also define an entropy H per second by dividing, not by n, but by the time T in seconds for n samples. Since $n = 2TW$, $H' = 2WH$.

With white thermal noise p is gaussian and we have

$$H' = \log \sqrt{2\pi eN},$$

$$H = W \log 2\pi eN.$$

For a given average power N, white noise has the maximum possible entropy. This follows from the maximizing properties of the Gaussian distribution noted above.

The entropy for a continuous stochastic process has many properties analogous to that for discrete processes. In the discrete case the entropy was related to the logarithm of the *probability* of long sequences, and to the *number* of reasonably probable sequences of long length. In the continuous case it is related in a similar fashion to the logarithm of the *probability density* for a long series of samples, and the *volume* of reasonably high probability in the function space.

More precisely, if we assume $p(x_1 \cdots x_n)$ continuous in all the x_i for all n, then for sufficiently large n

$$\left| \frac{\log p}{n} - H' \right| < \epsilon$$

for all choices of $(x_1, \cdots, x_n)$ apart from a set whose total probability is less than δ, with δ and ϵ arbitrarily small. This follows from the ergodic property if we divide the space into a large number of small cells.

The relation of H to volume can be stated as follows: Under the same assumptions consider the n dimensional space corresponding to $p(x_1, \cdots, x_n)$. Let $V_n(q)$ be the smallest volume in this space which includes in its interior a total probability q. Then

$$\lim_{n \to \infty} \frac{\log V_n(q)}{n} = H'$$

provided q does not equal 0 or 1.

These results show that for large n there is a rather well-defined volume (at least in the logarithmic sense) of high probability, and that within this volume the probability density is relatively uniform (again in the logarithmic sense).

In the white noise case the distribution function is given by

$$p(x_1 \cdots x_n) = \frac{1}{(2\pi N)^{n/2}} \exp - \frac{1}{2N} \Sigma x_i^2 .$$

Since this depends only on Σx_i^2 tbe surfaces of equal probability density are spheres and the entire distribution has spherical symmetry. The region of high probability is a sphere of radius $\sqrt{nN}$. As $n \to \infty$ the probability of being outside a sphere of radius $\sqrt{n(N + \epsilon)}$ approaches zero and $\frac{1}{n}$ times the logarithm of the volume of the sphere approaches $\log \sqrt{2\pi e N}$.

In the continuous case it is convenient to work not with the entropy H of an ensemble but with a derived quantity which we will call the entropy power. This is defined as the power in a white noise limited to the same band as the original ensemble and having the same entropy. In other words if H' is the entropy of an ensemble its entropy power is

$$N_1 = \frac{1}{2\pi e} \exp 2H'.$$

In the geometrical picture this amounts to measuring the high probability volume by the squared radius of a sphere having the same volume. Since white noise has the maximum entropy for a given power, the entropy power of any noise is less than or equal to its actual power.

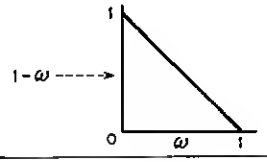
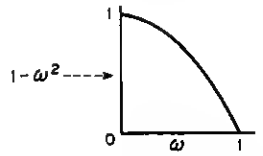
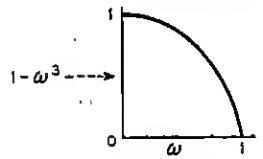
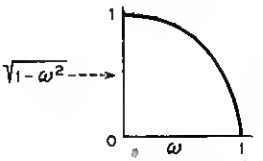
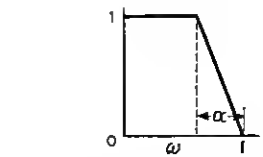
21. Entropy Loss in Linear Filters

Theorem 14: If an ensemble having an entropy H_1 per degree of freedom in band W is passed through a filter with characteristic $Y(f)$ the output ensemble has an entropy

$$H_2 = H_1 + \frac{1}{W} \int_W \log |Y(f)|^2 \, df.$$

The operation of the filter is essentially a linear transformation of coordinates. If we think of the different frequency components as the original coordinate system, the new frequency components are merely the old ones multiplied by factors. The coordinate transformation matrix is thus es-

TABLE I

GAIN	ENTROPY POWER FACTOR	ENTROPY POWER GAIN IN DECIBELS	IMPULSE RESPONSE
	$\frac{1}{e^2}$	-8.68	$\frac{\sin^2 \pi t}{(\pi t)^2}$
	$\left(\frac{2}{e}\right)^4$	-5.32	$2\left[\frac{\sin t}{t^3} - \frac{\cos t}{t^2}\right]$
	0.384	-4.15	$8\left[\frac{\cos t - 1}{t^4} - \frac{\cos t}{2t^2} + \frac{\sin t}{t^3}\right]$
	$\left(\frac{2}{e}\right)^2$	-2.66	$\frac{\pi}{2} \frac{J_1(t)}{t}$
	$\frac{1}{e^{2\alpha}}$	-8.68α	$\frac{1}{\alpha t^2}\left[\cos(1-\alpha)t - \cos t\right]$

sentially diagonalized in terms of these coordinates. The Jacohian of the transformation is (for n sine and n cosine components)

$$J = \prod_{i=1}^{n} |Y(f_i)|^2$$

where the f_i are equally spaced through the band W. This becomes in the limit

$$\exp \frac{1}{W} \int_W \log \mid Y(f) \mid^2 df.$$

Since J is constant its average value is this same quantity and applying the theorem on the change of entropy with a change of coordinates, the result follows. We may also phrase it in terms of the entropy power. Thus if the entropy power of the first ensemble is N_1 that of the second is

$$N_1 \exp \frac{1}{W} \int_W \log \mid Y(f) \mid^2 df.$$

The final entropy power is the initial entropy power multiplied by the geometric mean gain of the filter. If the gain is measured in *db*, then the output entropy power will be increased by the arithmetic mean *db* gain over W.

In Table I the entropy power loss has been calculated (and also expressed in *db*) for a number of ideal gain characteristics. The impulsive responses of these filters are also given for $W = 2\pi$, with phase assumed to be 0.

The entropy loss for many other cases can be obtained from these results. For example the entropy power factor $\frac{1}{e^2}$ for the first case also applies to any gain characteristic obtained from 1 — ω by a measure preserving transformation of the ω axis. In particular a linearly increasing gain $G(\omega) = \omega$, or a "saw tooth" characteristic between 0 and 1 have the same entropy loss. The reciprocal gain has the reciprocal factor. Thus $\frac{1}{\omega}$ has the factor e^2. Raising the gain to any power raises the factor to this power.

22. Entropy of the Sum of Two Ensembles

If we have two ensembles of functions $f_\alpha(t)$ and $g_\beta(t)$ we can form a new ensemble by "addition." Suppose the first ensemble has the probability density function $p(x_1, \cdots, x_n)$ and the second $q(x_1, \cdots, x_n)$. Then the density function for the sum is given by the convolution:

$$r(x_1, \cdots, x_n) = \int \cdots \int p(y_1, \cdots, y_n) \cdot q(x_1 - y_1, \cdots, x_n - y_n)\, dy_1, dy_2, \cdots, dy_n .$$

Physically this corresponds to adding the noises or signals represented by the original ensembles of functions.

The following result is derived in Appendix 6.

Theorem 15: Let the average power of two ensembles be N_1 and N_2 and let their entropy powers be $\bar{N}_1$ and $\bar{N}_2$. Then the entropy power of the sum, $\bar{N}_3$, is bounded by

$$\bar{N}_1 + \bar{N}_2 \leq \bar{N}_3 \leq N_1 + N_2 .$$

White Gaussian noise has the peculiar property that it can absorb any other noise or signal ensemble which may be added to it with a resultant entropy power approximately equal to the sum of the white noise power and the signal power (measured from the average signal value, which is normally zero), provided the signal power is small, in a certain sense, compared to the noise.

Consider the function space associated with these ensembles having n dimensions. The white noise corresponds to a spherical Gaussian distribution in this space. The signal ensemble corresponds to another probability distribution, not necessarily Gaussian or spherical. Let the second moments of this distribution about its center of gravity be a_{ij}. That is, if $p(x_1, \cdots, x_n)$ is the density distribution function

$$a_{ij} = \int \cdots \int p(x_i - \alpha_i)(x_j - \alpha_j)\, dx_1, \cdots, dx_n$$

where the α_i are the coordinates of the center of gravity. Now a_{ij} is a positive definite quadratic form, and we can rotate our coordinate system to align it with the principal directions of this form. a_{ij} is then reduced to diagonal form b_{ii}. We require that each b_{ii} be small compared to N, the squared radius of the spherical distribution.

In this case the convolution of the noise and signal produce a Gaussian distribution whose corresponding quadratic form is

$$N + b_{ii} .$$

The entropy power of this distribution is

$$[\Pi(N + b_{ii})]^{1/n}$$

or approximately

$$= [(N)^n + \Sigma b_{ii}(N)^{n-1}]^{1/n}$$

$$\doteq N + \frac{1}{n}\Sigma b_{ii} .$$

The last term is the signal power, while the first is the noise power.

PART IV: THE CONTINUOUS CHANNEL

23. The Capacity of a Continuous Channel

In a continuous channel the input or transmitted signals will be continuous functions of time $f(t)$ belonging to a certain set, and the output or received signals will be perturbed versions of these. We will consider only the case where both transmitted and received signals are limited to a certain band W. They can then be specified, for a time T, by $2TW$ numbers, and their statistical structure by finite dimensional distribution functions. Thus the statistics of the transmitted signal will be determined by

$$P(x_1, \cdots, x_n) = P(x)$$

and those of the noise by the conditional probability distribution

$$P_{x_1,\cdots,x_n}(y_1, \cdots, y_n) = P_x(y).$$

The rate of transmission of information for a continuous channel is defined in a way analogous to that for a discrete channel, namely

$$R = H(x) - H_y(x)$$

where $H(x)$ is the entropy of the input and $H_y(x)$ the equivocation. The channel capacity C is defined as the maximum of R when we vary the input over all possible ensembles. This means that in a finite dimensional approximation we must vary $P(x) = P(x_1, \cdots, x_n)$ and maximize

$$-\int P(x) \log P(x)\, dx + \int\int P(x, y) \log \frac{P(x, y)}{P(y)}\, dx\, dy.$$

This can be written

$$\int\int P(x, y) \log \frac{P(x, y)}{P(x)P(y)}\, dx\, dy$$

using the fact that $\int\int P(x, y) \log P(x)\, dx\, dy = \int P(x) \log P(x)\, dx$. The channel capacity is thus expressed

$$C = \operatorname*{Lim}_{T\to\infty} \operatorname*{Max}_{P(x)} \frac{1}{T} \int\int P(x, y) \log \frac{P(x, y)}{P(x)P(y)}\, dx\, dy.$$

It is obvious in this form that R and C are independent of the coordinate system since the numerator and denominator in $\log \frac{P(x, y)}{P(x)P(y)}$ will be multiplied by the same factors when x and y are transformed in any one to one way. This integral expression for C is more general than $H(x) - H_y(x)$. Properly interpreted (see Appendix 7) it will always exist while $H(x) - H_y(x)$

may assume an indeterminate form $\infty - \infty$ in some cases. This occurs, for example, if x is limited to a surface of fewer dimensions than n in its n dimensional approximation.

If the logarithmic base used in computing $H(x)$ and $H_y(x)$ is two then C is the maximum number of binary digits that can be sent per second over the channel with arbitrarily small equivocation, just as in the discrete case. This can be seen physically by dividing the space of signals into a large number of small cells, sufficiently small so that the probability density $P_x(y)$ of signal x being perturbed to point y is substantially constant over a cell (either of x or y). If the cells are considered as distinct points the situation is essentially the same as a discrete channel and the proofs used there will apply. But it is clear physically that this quantizing of the volume into individual points cannot in any practical situation alter the final answer significantly, provided the regions are sufficiently small. Thus the capacity will be the limit of the capacities for the discrete subdivisions and this is just the continuous capacity defined above.

On the mathematical side it can be shown first (see Appendix 7) that if u is the message, x is the signal, y is the received signal (perturbed by noise) and v the recovered message then

$$H(x) - H_y(x) \geq H(u) - H_v(u)$$

regardless of what operations are performed on u to obtain x or on y to obtain v. Thus no matter how we encode the binary digits to obtain the signal, or how we decode the received signal to recover the message, the discrete rate for the binary digits does not exceed the channel capacity we have defined. On the other hand, it is possible under very general conditions to find a coding system for transmitting binary digits at the rate C with as small an equivocation or frequency of errors as desired. This is true, for example, if, when we take a finite dimensional approximating space for the signal functions, $P(x, y)$ is continuous in both x and y except at a set of points of probability zero.

An important special case occurs when the noise is added to the signal and is independent of it (in the probability sense). Then $P_x(y)$ is a function only of the difference $n = (y - x)$,

$$P_x(y) = Q(y - x)$$

and we can assign a definite entropy to the noise (independent of the statistics of the signal), namely the entropy of the distribution $Q(n)$. This entropy will be denoted by $H(n)$.

Theorem 16: If the signal and noise are independent and the received signal is the sum of the transmitted signal and the noise then the rate of

transmission is

$$R = H(y) - H(n)$$

i.e., the entropy of the received signal less the entropy of the noise. The channel capacity is

$$C = \operatorname*{Max}_{P(x)} H(y) - H(n).$$

We have, since $y = x + n$:

$$H(x, y) = H(x, n).$$

Expanding the left side and using the fact that x and n are independent

$$H(y) + H_y(x) = H(x) + H(n).$$

Hence

$$R = H(x) - H_y(x) = H(y) - H(n).$$

Since $H(n)$ is independent of $P(x)$, maximizing R requires maximizing $H(y)$, the entropy of the received signal. If there are certain constraints on the ensemble of transmitted signals, the entropy of the received signal must be maximized subject to these constraints.

24. Channel Capacity with an Average Power Limitation

A simple application of Theorem 16 is the case where the noise is a white thermal noise and the transmitted signals are limited to a certain average power P. Then the received signals have an average power $P + N$ where N is the average noise power. The maximum entropy for the received signals occurs when they also form a white noise ensemble since this is the greatest possible entropy for a power $P + N$ and can be obtained by a suitable choice of the ensemble of transmitted signals, namely if they form a white noise ensemble of power P. The entropy (per second) of the received ensemble is then

$$H(y) = W \log 2\pi e(P + N),$$

and the noise entropy is

$$H(n) = W \log 2\pi eN.$$

The channel capacity is

$$C = H(y) - H(n) = W \log \frac{P + N}{N}.$$

Summarizing we have the following:

Theorem 17: The capacity of a channel of band W perturbed by white

thermal noise of power N when the average transmitter power is P is given by

$$C = W \log \frac{P + N}{N}.$$

This means of course that by sufficiently involved encoding systems we can transmit binary digits at the rate $W \log_2 \frac{P + N}{N}$ bits per second, with arbitrarily small frequency of errors. It is not possible to transmit at a higher rate by any encoding system without a definite positive frequency of errors.

To approximate this limiting rate of transmission the transmitted signals must approximate, in statistical properties, a white noise.[6] A system which approaches the ideal rate may be described as follows: Let $M = 2^s$ samples of white noise be constructed each of duration T. These are assigned binary numbers from 0 to $(M - 1)$. At the transmitter the message sequences are broken up into groups of s and for each group the corresponding noise sample is transmitted as the signal. At the receiver the M samples are known and the actual received signal (perturbed by noise) is compared with each of them. The sample which has the least R.M.S. discrepancy from the received signal is chosen as the transmitted signal and the corresponding binary number reconstructed. This process amounts to choosing the most probable (*a posteriori*) signal. The number M of noise samples used will depend on the tolerable frequency ϵ of errors, but for almost all selections of samples we have

$$\operatorname*{Lim}_{\epsilon \to 0} \operatorname*{Lim}_{T \to \infty} \frac{\log M(\epsilon, T)}{T} = W \log \frac{P + N}{N},$$

so that no matter how small ϵ is chosen, we can, by taking T sufficiently large, transmit as near as we wish to $TW \log \frac{P + N}{N}$ binary digits in the time T.

Formulas similar to $C = W \log \frac{P + N}{N}$ for the white noise case have been developed independently by several other writers, although with somewhat different interpretations. We may mention the work of N. Wiener,[7] W. G. Tuller,[8] and H. Sullivan in this connection.

In the case of an arbitrary perturbing noise (not necessarily white thermal noise) it does not appear that the maximizing problem involved in deter-

[6] This and other properties of the white noise case are discussed from the geometrical point of view in "Communication in the Presence of Noise," loc. cit.

[7] "Cybernetics," loc. cit.

[8] Sc. D. thesis, Department of Electrical Engineering, M.I.T., 1948.

mining the channel capacity C can be solved explicitly. However, upper and lower bounds can be set for C in terms of the average noise power N and the noise entropy power N_1. These bounds are sufficiently close together in most practical cases to furnish a satisfactory solution to the problem.

Theorem 18: The capacity of a channel of band W perturbed by an arbitrary noise is bounded by the inequalities

$$W \log \frac{P + N_1}{N_1} \leq C \leq W \log \frac{P + N}{N_1}$$

where

P = average transmitter power
N = average noise power
N_1 = entropy power of the noise.

Here again the average power of the perturbed signals will be $P + N$. The maximum entropy for this power would occur if the received signal were white noise and would be $W \log 2\pi e(P + N)$. It may not be possible to achieve this; i.e. there may not be any ensemble of transmitted signals which, added to the perturbing noise, produce a white thermal noise at the receiver, but at least this sets an upper bound to $H(y)$. We have, therefore

$$\begin{aligned} C &= \max H(y) - H(n) \\ &\leq W \log 2\pi e(P + N) - W \log 2\pi e N_1 . \end{aligned}$$

This is the upper limit given in the theorem. The lower limit can be obtained by considering the rate if we make the transmitted signal a white noise, of power P. In this case the entropy power of the received signal must be at least as great as that of a white noise of power $P + N_1$ since we have shown in a previous theorem that the entropy power of the sum of two ensembles is greater than or equal to the sum of the individual entropy powers. Hence

$$\max H(y) \geq W \log 2\pi e(P + N_1)$$

and

$$\begin{aligned} C &\geq W \log 2\pi e(P + N_1) - W \log 2\pi e N_1 \\ &= W \log \frac{P + N_1}{N_1} . \end{aligned}$$

As P increases, the upper and lower bounds approach each other, so we have as an asymptotic rate

$$W \log \frac{P + N}{N_1}$$

If the noise is itself white, $N = N_1$ and the result reduces to the formula proved previously:

$$C = W \log \left(1 + \frac{P}{N}\right).$$

If the noise is Gaussian but with a spectrum which is not necessarily flat, N_1 is the geometric mean of the noise power over the various frequencies in the band W. Thus

$$N_1 = \exp \frac{1}{W} \int_W \log N(f)\, df$$

where $N(f)$ is the noise power at frequency f.

Theorem 19: If we set the capacity for a given transmitter power P equal to

$$C = W \log \frac{P + N - \eta}{N_1}$$

then η is monotonic decreasing as P increases and approaches 0 as a limit.

Suppose that for a given power P_1 the channel capacity is

$$W \log \frac{P_1 + N - \eta_1}{N_1}$$

This means that the best signal distribution, say $p(x)$, when added to the noise distribution $q(x)$, gives a received distribution $r(y)$ whose entropy power is $(P_1 + N - \eta_1)$. Let us increase the power to $P_1 + \Delta P$ by adding a white noise of power ΔP to the signal. The entropy of the received signal is now at least

$$H(y) = W \log 2\pi e(P_1 + N - \eta_1 + \Delta P)$$

by application of the theorem on the minimum entropy power of a sum. Hence, since we can attain the H indicated, the entropy of the maximizing distribution must be at least as great and η must be monotonic decreasing. To show that $\eta \to 0$ as $P \to \infty$ consider a signal which is a white noise with a large P. Whatever the perturbing noise, the received signal will be approximately a white noise, if P is sufficiently large, in the sense of having an entropy power approaching $P + N$.

25. The Channel Capacity with a Peak Power Limitation

In some applications the transmitter is limited not by the average power output but by the peak instantaneous power. The problem of calculating the channel capacity is then that of maximizing (by variation of the ensemble of transmitted symbols)

$$H(y) - H(n)$$

subject to the constraint that all the functions $f(t)$ in the ensemble be less than or equal to $\sqrt{S}$, say, for all t. A constraint of this type does not work out as well mathematically as the average power limitation. The most we have obtained for this case is a lower bound valid for all $\frac{S}{N}$, an "asymptotic" upper band $\left(\text{valid for large } \frac{S}{N}\right)$ and an asymptotic value of C for $\frac{S}{N}$ small.

Theorem 20: The channel capacity C for a band W perturbed by white thermal noise of power N is bounded by

$$C \geq W \log \frac{2}{\pi \epsilon^3} \frac{S}{N},$$

where S is the peak allowed transmitter power. For sufficiently large $\frac{S}{N}$

$$C \leq W \log \frac{\frac{2}{\pi e} S + N}{N} (1 + \epsilon)$$

where ϵ is arbitrarily small. As $\frac{S}{N} \to 0$ (and provided the band W starts at 0)

$$C \to W \log \left(1 + \frac{S}{N}\right).$$

We wish to maximize the entropy of the received signal. If $\frac{S}{N}$ is large this will occur very nearly when we maximize the entropy of the transmitted ensemble.

The asymptotic upper bound is obtained by relaxing the conditions on the ensemble. Let us suppose that the power is limited to S not at every instant of time, but only at the sample points. The maximum entropy of the transmitted ensemble under these weakened conditions is certainly greater than or equal to that under the original conditions. This altered problem can be solved easily. The maximum entropy occurs if the different samples are independent and have a distribution function which is constant from $-\sqrt{S}$ to $+\sqrt{S}$. The entropy can be calculated as

$$W \log 4S.$$

The received signal will then have an entropy less than

$$W \log (4S + 2\pi eN)(1 + \epsilon)$$

with $\epsilon \to 0$ as $\frac{S}{N} \to \infty$ and the channel capacity is obtained by subtracting tbe entropy of the white noise, $W \log 2\pi eN$

$$W \log (4S + 2\pi eN)(1 + \epsilon) - W \log (2\pi eN) = W \log \frac{\frac{2}{\pi e} S + N}{N} (1 + \epsilon).$$

This is the desired upper hound to the channel capacity.

To obtain a lower bound consider the same ensemble of functions. Let these functions be passed through an ideal filter with a triangular transfer characteristic. The gain is to be unity at frequency 0 and decline linearly down to gain 0 at frequency W. We first show that the output functions of the filter have a peak power limitation S at all times (not just the sample points). First we note that a pulse $\frac{\sin 2\pi Wt}{2\pi Wt}$ going into the filter produces

$$\frac{1}{2} \frac{\sin^2 \pi Wt}{(\pi Wt)^2}$$

in the output. This function is never negative. The input function (in the general case) can be thought of as the sum of a series of shifted functions

$$a \frac{\sin 2\pi Wt}{2\pi Wt}$$

where a, the amplitude of the sample, is not greater than $\sqrt{S}$. Hence the output is the sum of shifted functions of the non-negative form above witb the same coefficients. These functions heing non-negative, the greatest positive value for any t is ohtained when all the coefficients a have their maximum positive values, i.e. $\sqrt{S}$. In this case the input function was a constant of amplitude $\sqrt{S}$ and since the filter has unit gain for D.C., the output is the same. Hence the output ensemhle has a peak power S.

The entropy of the output ensemhle can be calculated from that of the input ensemhle by using the theorem dealing with such a situation. The output entropy is equal to the input entropy plus the geometrical mean gain of the filter;

$$\int_0^W \log G^2 \, df = \int_0^W \log \left(\frac{W - f}{W} \right)^2 df = -2W$$

Hence the output entropy is

$$W \log 4S - 2W = W \log \frac{4S}{e^2}$$

and the channel capacity is greater than

$$W \log \frac{2}{\pi e^3} \frac{S}{N}.$$

We now wish to show that, for small $\frac{S}{N}$ (peak signal power over average white noise power), the channel capacity is approximately

$$C = W \log \left(1 + \frac{S}{N}\right).$$

More precisely $C/W \log \left(1 + \frac{S}{N}\right) \to 1$ as $\frac{S}{N} \to 0$. Since the average signal power P is less than or equal to the peak S, it follows that for all $\frac{S}{N}$

$$C \leq W \log \left(1 + \frac{P}{N}\right) \leq W \log \left(1 + \frac{S}{N}\right).$$

Therefore, if we can find an ensemble of functions such that they correspond to a rate nearly $W \log \left(1 + \frac{S}{N}\right)$ and are limited to band W and peak S the result will be proved. Consider the ensemble of functions of the following type. A series of t samples have the same value, either $+\sqrt{S}$ or $-\sqrt{S}$, then the next t samples have the same value, etc. The value for a series is chosen at random, probability $\frac{1}{2}$ for $+\sqrt{S}$ and $\frac{1}{2}$ for $-\sqrt{S}$. If this ensemble be passed through a filter with triangular gain characteristic (unit gain at D.C.), the output is peak limited to $\pm S$. Furthermore the average power is nearly S and can be made to approach this by taking t sufficiently large. The entropy of the sum of this and the thermal noise can be found by applying the theorem on the sum of a noise and a small signal. This theorem will apply if

$$\sqrt{t} \frac{S}{N}$$

is sufficiently small. This can be insured by taking $\frac{S}{N}$ small enough (after t is chosen). The entropy power will be $S + N$ to as close an approximation as desired, and hence the rate of transmission as near as we wish to

$$W \log \left(\frac{S + N}{N}\right).$$

PART V: THE RATE FOR A CONTINUOUS SOURCE

26. Fidelity Evaluation Functions

In the case of a discrete source of information we were able to determine a definite rate of generating information, namely the entropy of the underlying stochastic process. With a continuous source the situation is considerably more involved. In the first place a continuously variable quantity can assume an infinite number of values and requires, therefore, an infinite number of binary digits for exact specification. This means that to transmit the output of a continuous source with *exact recovery* at the receiving point requires, in general, a channel of infinite capacity (in bits per second). Since, ordinarily, channels have a certain amount of noise, and therefore a finite capacity, exact transmission is impossible.

This, however, evades the real issue. Practically, we are not interested in exact transmission when we have a continuous source, but only in transmission to within a certain tolerance. The question is, can we assign a definite rate to a continuous source when we require only a certain fidelity of recovery, measured in a suitable way. Of course, as the fidelity requirements are increased the rate will increase. It will be shown that we can, in very general cases, define such a rate, having the property that it is possible, by properly encoding the information, to transmit it over a channel whose capacity is equal to the rate in question, and satisfy the fidelity requirements. A channel of smaller capacity is insufficient.

It is first necessary to give a general mathematical formulation of the idea of fidelity of transmission. Consider the set of messages of a long duration, say T seconds. The source is described by giving the probability density, in the associated space, that the source will select the message in question $P(x)$. A given communication system is described (from the external point of view) by giving the conditional probability $P_x(y)$ that if message x is produced by the source the recovered message at the receiving point will be y. The system as a whole (including source and transmission system) is described by the probability function $P(x, y)$ of having message x and final output y. If this function is known, the complete characteristics of the system from the point of view of fidelity are known. Any evaluation of fidelity must correspond mathematically to an operation applied to $P(x, y)$. This operation must at least have the properties of a simple ordering of systems; i.e. it must be possible to say of two systems represented by $P_1(x, y)$ and $P_2(x, y)$ that, according to our fidelity criterion, either (1) the first has higher fidelity, (2) the second has higher fidelity, or (3) they have

equal fidelity. This means that a criterion of fidelity can be represented by a numerically valued function:

$$v(P(x, y))$$

whose argument ranges over possible probability functions $P(x, y)$.

We will now show that under very general and reasonable assumptions the function $v(P(x, y))$ can be written in a seemingly much more specialized form, namely as an average of a function $\rho(x, y)$ over the set of possible values of x and y:

$$v(P(x, y)) = \int\int P(x, y)\, \rho(x, y)\, dx\, dy$$

To obtain this we need only assume (1) that the source and system are ergodic so that a very long sample will be, with probability nearly 1, typical of the ensemble, and (2) that the evaluation is "reasonable" in the sense that it is possible, by observing a typical input and output x_1 and y_1, to form a tentative evaluation on the basis of these samples; and if these samples are increased in duration the tentative evaluation will, with probability 1, approach the exact evaluation based on a full knowledge of $P(x, y)$. Let the tentative evaluation be $\rho(x, y)$. Then the function $\rho(x, y)$ approaches (as $T \to \infty$) a constant for almost all (x, y) which are in the high probability region corresponding to the system:

$$\rho(x, y) \to v(P(x, y))$$

and we may also write

$$\rho(x, y) \to \int\int P(x, y)\rho(x, y)\, dx, dy$$

since

$$\int\int P(x, y)\, dx\, dy = 1$$

This establishes the desired result.

The function $\rho(x, y)$ has the general nature of a "distance" between x and y.[9] It measures how bad it is (according to our fidelity criterion) to receive y when x is transmitted. The general result given above can be restated as follows: Any reasonable evaluation can be represented as an average of a distance function over the set of messages and recovered messages x and y weighted according to the probability $P(x, y)$ of getting the pair in question, provided the duration T of the messages be taken sufficiently large.

[9] It is not a "metric" in the strict sense, however, since in general it does not satisfy either $\rho(x, y) = \rho(y, x)$ or $\rho(x, y) + \rho(y, z) \geq \rho(x, z)$.

The following are simple examples of evaluation functions:

1. R.M.S. Criterion.

$$v = \overline{(x(t) - y(t))^2}$$

In this very commonly used criterion of fidelity the distance function $\rho(x, y)$ is (apart from a constant factor) the square of the ordinary euclidean distance between the points x and y in the associated function space.

$$\rho(x, y) = \frac{1}{T}\int_0^T [x(t) - y(t)]^2\, dt$$

2. Frequency weighted R.M.S. criterion. More generally one can apply different weights to the different frequency components before using an R.M.S. measure of fidelity. This is equivalent to passing the difference $x(t) - y(t)$ through a shaping filter and then determining the average power in the output. Thus let

$$e(t) = x(t) - y(t)$$

and

$$f(t) = \int_{-\infty}^{\infty} e(\tau)k(t - \tau)\, dt$$

then

$$\rho(x, y) = \frac{1}{T}\int_0^T f(t)^2\, dt.$$

3. Absolute error criterion.

$$\rho(x, y) = \frac{1}{T}\int_0^T |\, x(t) - y(t)\,|\, dt$$

4. The structure of the ear and brain determine implicitly an evaluation, or rather a number of evaluations, appropriate in the case of speech or music transmission. There is, for example, an "intelligibility" criterion in which $\rho(x, y)$ is equal to the relative frequency of incorrectly interpreted words when message $x(t)$ is received as $y(t)$. Although we cannot give an explicit representation of $\rho(x, y)$ in these cases it could, in principle, be determined by sufficient experimentation. Some of its properties follow from well-known experimental results in hearing, e.g., the ear is relatively insensitive to phase and the sensitivity to amplitude and frequency is roughly logarithmic.
5. The discrete case can be considered as a specialization in which we have

tacitly assumed an evaluation based on the frequency of errors. The function $\rho(x, y)$ is then defined as the number of symbols in the sequence y differing from the corresponding symbols in x divided by the total number of symbols in x.

27. The Rate for a Source Relative to a Fidelity Evaluation

We are now in a position to define a rate of generating information for a continuous source. We are given $P(x)$ for the source and an evaluation v determined by a distance function $\rho(x, y)$ which will be assumed continuous in both x and y. With a particular system $P(x, y)$ the quality is measured by

$$v = \iint \rho(x, y)\ P(x, y)\ dx\ dy$$

Furthermore the rate of flow of binary digits corresponding to $P(x, y)$ is

$$R = \iint P(x, y)\ \log \frac{P(x, y)}{P(x)P(y)}\ dx\ dy.$$

We define the rate R_1 of generating information for a given quality v_1 of reproduction to be the minimum of R when we keep v fixed at v_1 and vary $P_x(y)$. That is:

$$R_1 = \operatorname*{Min}_{P_x(y)} \iint P(x, y)\ \log \frac{P(x, y)}{P(x)P(y)}\ dx\ dy$$

subject to the constraint:

$$v_1 = \iint P(x, y)\rho(x, y)\ dx\ dy.$$

This means that we consider, in effect, all the communication systems that might be used and that transmit with the required fidelity. The rate of transmission in bits per second is calculated for each one and we choose that having the least rate. This latter rate is the rate we assign the source for the fidelity in question.

The justification of this definition lies in the following result:

Theorem 21: If a source has a rate R_1 for a valuation v_1 it is possible to encode the output of the source and transmit it over a channel of capacity C with fidelity as near v_1 as desired provided $R_1 \leq C$. This is not possible if $R_1 > C$.

The last statement in the theorem follows immediately from the definition of R_1 and previous results. If it were not true we could transmit more than C bits per second over a channel of capacity C. The first part of the theorem is proved by a method analogous to that used for Theorem 11. We may, in the first place, divide the (x, y) space into a large number of small cells and

represent the situation as a discrete case. This will not change the evaluation function by more than an arbitrarily small amount (when the cells are very small) because of the continuity assumed for $p(x, y)$. Suppose that $P_1(x, y)$ is the particular system which minimizes the rate and gives R_1. We choose from the high probability y's a set at random containing

$$2^{(R_1+\epsilon)T}$$

members where $\epsilon \to 0$ as $T \to \infty$. With large T each chosen point will be connected by a high probability line (as in Fig. 10) to a set of x's. A calculation similar to that used in proving Theorem 11 shows that with large T almost all x's are covered by the fans from the chosen y points for almost all choices of the y's. The communication system to be used operates as follows: The selected points are assigned binary numbers. When a message x is originated it will (with probability approaching 1 as $T \to \infty$) lie within one at least of the fans. The corresponding binary number is transmitted (or one of them chosen arbitrarily if there are several) over the channel by suitable coding means to give a small probability of error. Since $R_1 \leq C$ this is possible. At the receiving point the corresponding y is reconstructed and used as the recovered message.

The evaluation v_1' for this system can be made arbitrarily close to v_1 by taking T sufficiently large. This is due to the fact that for each long sample of message $x(t)$ and recovered message $y(t)$ the evaluation approaches v_1 (with probability 1).

It is interesting to note that, in this system, the noise in the recovered message is actually produced by a kind of general quantizing at the transmitter and is not produced by the noise in the channel. It is more or less analogous to the quantizing noise in P.C.M.

28. The Calculation of Rates

The definition of the rate is similar in many respects to the definition of channel capacity. In the former

$$R = \operatorname*{Max}_{P_x(y)} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)}\, dx\, dy$$

with $P(x)$ and $v_1 = \iint P(x, y)\rho(x, y)\, dx\, dy$ fixed. In the latter

$$C = \operatorname*{Min}_{P(x)} \iint P(x, y) \log \frac{P(x, y)}{P(x)P(y)}\, dx\, dy$$

with $P_x(y)$ fixed and possibly one or more other constraints (e.g., an average power limitation) of the form $K = \int\int P(x, y)\, \lambda(x, y)\, dx\, dy$.

A partial solution of the general maximizing problem for determining the rate of a source can be given. Using Lagrange's method we consider

$$\iint \left[P(x, y) \log \frac{P(x, y)}{P(x)P(y)} + \mu P(x, y)\rho(x, y) + \nu(x)P(x, y) \right] dx\, dy$$

The variational equation (when we take the first variation on $P(x, y)$) leads to

$$P_y(x) = B(x)\, e^{-\lambda \rho(x,y)}$$

where λ is determined to give the required fidelity and $B(x)$ is chosen to satisfy

$$\int B(x) e^{-\lambda \rho(x,y)}\, dx = 1$$

This shows that, with best encoding, the conditional probability of a certain cause for various received y, $P_y(x)$ will decline exponentially with the distance function $\rho(x, y)$ between the x and y is question.

In the special case where the distance function $\rho(x, y)$ depends only on the (vector) difference between x and y,

$$\rho(x, y) = \rho(x - y)$$

we have

$$\int B(x) e^{-\lambda \rho(x-y)}\, dx = 1.$$

Hence $B(x)$ is constant, say α, and

$$P_y(x) = \alpha e^{-\lambda \rho(x-y)}$$

Unfortunately these formal solutions are difficult to evaluate in particular cases and seem to be of little value. In fact, the actual calculation of rates has been carried out in only a few very simple cases.

If the distance function $\rho(x, y)$ is the mean square discrepancy between x and y and the message ensemble is white noise, the rate can be determined. In that case we have

$$R = \text{Min } [H(x) - H_y(x)] = H(x) - \text{Max } H_y(x)$$

with $N = \overline{(x - y)^2}$. But the Max $H_y(x)$ occurs when $y - x$ is a white noise, and is equal to $W_1 \log 2\pi e N$ where W_1 is the bandwidth of the message ensemble. Therefore

$$\begin{aligned} R &= W_1 \log 2\pi e Q - W_1 \log 2\pi e N \\ &= W_1 \log \frac{Q}{N} \end{aligned}$$

where Q is the average message power. This proves the following:

Theorem 22: The rate for a white noise source of power Q and band W_1 relative to an R.M.S. measure of fidelity is

$$R = W_1 \log \frac{Q}{N}$$

where N is the allowed mean square error between original and recovered messages.

More generally with any message source we can obtain inequalities bounding the rate relative to a mean square error criterion.

Theorem 23: The rate for any source of band W_1 is bounded by

$$W_1 \log \frac{Q_1}{N} \leq R \leq W_1 \log \frac{Q}{N}$$

where Q is the average power of the source, Q_1 its entropy power and N the allowed mean square error.

The lower bound follows from the fact that the max $H_y(x)$ for a given $(x - y)^2 = N$ occurs in the white noise case. The upper bound results if we place the points (used in the proof of Theorem 21) not in the best way but at random in a sphere of radius $\sqrt{Q - N}$.

Acknowledgments

The writer is indebted to his colleagues at the Laboratories, particularly to Dr. H. W. Bode, Dr. J. R. Pierce, Dr. B. McMillan, and Dr. B. M. Oliver for many helpful suggestions and criticisms during the course of this work. Credit should also be given to Professor N. Wiener, whose elegant solution of the problems of filtering and prediction of stationary ensembles has considerably influenced the writer's thinking in this field.

APPENDIX 5

Let S_1 be any measurable subset of the g ensemble, and S_2 the subset of the f ensemble which gives S_1 under the operation T. Then

$$S_1 = TS_2.$$

Let H^λ be the operator which shifts all functions in a set by the time λ. Then

$$H^\lambda S_1 = H^\lambda T S_2 = T H^\lambda S_2$$

since T is invariant and therefore commutes with H^λ. Hence if $m[S]$ is the probability measure of the set S

$$m[H^\lambda S_1] = m[T H^\lambda S_2] = m[H^\lambda S_2]$$
$$= m[S_2] = m[S_1]$$

where the second equality is by definition of measure in the g space the third since the f ensemble is stationary, and the last by definition of g measure again.

To prove that the ergodic property is preserved under invariant operations, let S_1 be a subset of the g ensemble which is invariant under H^λ, and let S_2 be the set of all functions f which transform into S_1. Then

$$H^\lambda S_1 = H^\lambda T S_2 = T H^\lambda S_2 = S_1$$

so that $H^\lambda S_1$ is included in S_1 for all λ. Now, since

$$m[H^\lambda S_2] = m[S_1]$$

this implies

$$H^\lambda S_2 = S_2$$

for all λ with $m[S_2] \neq 0, 1$. This contradiction shows that S_1 does not exist.

APPENDIX 6

The upper bound, $\bar{N}_3 \leq N_1 + N_2$, is due to the fact that the maximum possible entropy for a power $N_1 + N_2$ occurs when we have a white noise of this power. In this case the entropy power is $N_1 + N_2$.

To obtain the lower bound, suppose we have two distributions in n dimensions $p(x_i)$ and $q(x_i)$ with entropy powers $\bar{N}_1$ and $\bar{N}_2$. What form should p and q have to minimize the entropy power $\bar{N}_3$ of their convolution $r(x_i)$:

$$r(x_i) = \int p(y_i) q(x_i - y_i)\, dy_i .$$

The entropy H_3 of r is given by

$$H_3 = -\int r(x_i) \log r(x_i)\, dx_i .$$

We wish to minimize this subject to the constraints

$$H_1 = -\int p(x_i) \log p(x_i)\, dx_i$$

$$H_2 = -\int q(x_i) \log q(x_i)\, dx_i .$$

We consider then

$$U = -\int [r(x) \log r(x) + \lambda p(x) \log p(x) + \mu q(x) \log q(x)]\, dx$$

$$\delta U = -\int [[1 + \log r(x)]\delta r(x) + \lambda[1 + \log p(x)]\delta p(x) + \mu[1 + \log q(x)\delta q(x)]]\, dx.$$

If $p(x)$ is varied at a particular argument $x_i = s_i$, the variation in $r(x)$ is

$$\delta r(x) = q(x_i - s_i)$$

and

$$\delta U = -\int q(x_i - s_i) \log r(x_i)\, dx_i - \lambda \log p(s_i) = 0$$

and similarly when q is varied. Hence the conditions for a minimum are

$$\int q(x_i - s_i) \log r(x_i) = -\lambda \log p(s_i)$$

$$\int p(x_i - s_i) \log r(x_i) = -\mu \log q(s_i).$$

If we multiply the first by $p(s_i)$ and the second by $q(s_i)$ and integrate with respect to s we obtain

$$H_3 = -\lambda H_1$$

$$H_3 = -\mu H_2$$

or solving for λ and μ and replacing in the equations

$$H_1 \int q(x_i - s_i) \log r(x_i)\, dx_i = -H_3 \log p(s_i)$$

$$H_2 \int p(x_i - s_i) \log r(x_i)\, dx_i = -H_3 \log p(s_i).$$

Now suppose $p(x_i)$ and $q(x_i)$ are normal

$$p(x_i) = \frac{|A_{ij}|^{n/2}}{(2\pi)^{n/2}} \exp -\tfrac{1}{2}\Sigma A_{ij} x_i x_j$$

$$q(x_i) = \frac{|B_{ij}|^{n/2}}{(2\pi)^{n/2}} \exp -\tfrac{1}{2}\Sigma B_{ij} x_i x_j .$$

Then $r(x_i)$ will also be normal with quadratic form C_{ij}. If the inverses of these forms are a_{ij}, b_{ij}, c_{ij} then

$$c_{ij} = a_{ij} + b_{ij} .$$

We wish to show that these functions satisfy the minimizing conditions if and only if $a_{ij} = Kb_{ij}$ and thus give the minimum H_3 under the constraints. First we have

$$\log r(x_i) = \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \tfrac{1}{2}\Sigma C_{ij} x_i x_j$$

$$\int q(x_i - s_i) \log r(x_i) = \frac{n}{2} \log \frac{1}{2\pi} |C_{ij}| - \tfrac{1}{2}\Sigma C_{ij} s_i s_j - \tfrac{1}{2}\Sigma C_{ij} b_{ij} .$$

This should equal

$$\frac{H_3}{H_1}\left[\frac{n}{2}\log\frac{1}{2\pi}|A_{ij}| - \tfrac{1}{2}\Sigma A_{ij}s_i s_j\right]$$

which requires $A_{ij} = \dfrac{H_1}{H_3}C_{ij}$.

In this case $A_{ij} = \dfrac{H_1}{H_2}B_{ij}$ and both equations reduce to identities.

APPENDIX 7

The following will indicate a more general and more rigorous approach to the central definitions of communication theory. Consider a probability measure space whose elements are ordered pairs (x, y). The variables x, y are to be identified as the possible transmitted and received signals of some long duration T. Let us call the set of all points whose x belongs to a subset S_1 of x points the strip over S_1, and similarly the set whose y belongs to S_2 the strip over S_2. We divide x and y into a collection of non-overlapping measurable subsets X_i and Y_i approximate to the rate of transmission R by

$$R_1 = \frac{1}{T}\sum_i P(X_i, Y_i)\log\frac{P(X_i, Y_i)}{P(X_i)P(Y_i)}$$

where

$P(X_i)$ is the probability measure of the strip over X_i
$P(Y_i)$ is the probability measure of the strip over Y_i
$P(X_i, Y_i)$ is the probability measure of the intersection of the strips.

A further subdivision can never decrease R_1. For let X_1 be divided into $X_1 = X_1' + X_1''$ and let

$$P(Y_1) = a \qquad P(X_1) = b + c$$
$$P(X_1') = b \qquad P(X_1', Y_1) = d$$
$$P(X_1'') = c \qquad P(X_1'', Y_1) = e$$
$$P(X_1, Y_1) = d + e$$

Then in the sum we have replaced (for the X_1, Y_1 intersection)

$$(d + e)\log\frac{d + e}{a(b + c)} \quad \text{by} \quad d\log\frac{d}{ab} + e\log\frac{e}{ac}.$$

It is easily shown that with the limitation we have on b, c, d, e,

$$\left[\frac{d + e}{b + c}\right]^{d+e} \leq \frac{d^d e^e}{b^d c^e}$$

and consequently the sum is increased. Thus the various possible subdivisions form a directed set, with R monotonic increasing with refinement of the subdivision. We may define R unambiguously as the least upper bound for the R_1 and write it

$$R = \frac{1}{T} \int\int P(x, y) \log \frac{P(x, y)}{P(x)P(y)} \, dx \, dy.$$

This integral, understood in the above sense, includes both the continuous and discrete cases and of course many others which cannot be represented in either form. It is trivial in this formulation that if x and u are in one-to-one correspondence, the rate from u to y is equal to that from x to y. If v is any function of y (not necessarily with an inverse) then the rate from x to y is greater than or equal to that from x to v since, in the calculation of the approximations, the subdivisions of y are essentially a finer subdivision of those for v. More generally if y and v are related not functionally but statistically, i.e., we have a probability measure space (y, v), then $R(x, v) \leq R(x, y)$. This means that any operation applied to the received signal, even though it involves statistical elements, does not increase R.

Another notion which should be defined precisely in an abstract formulation of the theory is that of "dimension rate," that is the average number of dimensions required per second to specify a member of an ensemble. In the band limited case 2W numbers per second are sufficient. A general definition can be framed as follows. Let $f_\alpha(t)$ be an ensemble of functions and let $\rho_T[f_\alpha(t), f_\beta(t)]$ be a metric measuring the "distance" from f_α to f_β over the time T (for example the R.M.S. discrepancy over this interval.) Let $N(\epsilon, \delta, T)$ be the least number of elements f which can be chosen such that all elements of the ensemble apart from a set of measure δ are within the distance ϵ of at least one of those chosen. Thus we are covering the space to within ϵ apart from a set of small measure δ. We define the dimension rate λ for the ensemble by the triple limit

$$\lambda = \lim_{\delta \to 0} \lim_{\epsilon \to 0} \lim_{T \to \infty} \frac{\log N(\epsilon, \delta, T)}{T \log \epsilon}.$$

This is a generalization of the measure type definitions of dimension in topology, and agrees with the intuitive dimension rate for simple ensembles where the desired result is obvious.